On the Optimal Error Rate of Stochastic Block Model with Symmetric Side Information

Feng Zhao[1], **Jin Sima**[2], and **Shao-Lun Huang**[3]

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China 100084
[2]Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA
[3]DSIT Research Center, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China 518055

***Abstract*—Side information improves the accuracy in community detection problems. While experimental results demonstrate the superior performance of many detection methods based on both the node attributes and graph structure, the question of the fundamental limit of the error rate for exact recovery remains open. In this paper, we obtain the asymptotic optimal error rate in the sense of exact recovery for a special two-community symmetric stochastic block model (SSBM) with side information consisting of multiple features. Our result provides insight on the number of features and nodes in the graph needed for community detection.**

I. Introduction

In network analysis, community detection assigns discrete labels to each node of the graph based on the observation of graph edges. In addition to the edge information, extra node features are often available in real-world applications in the form of graph signal [1], noisy labels [2], or feature vectors [3]. Combining the edge and node information, it is expected that better accuracy can be achieved for community detection problems. Within this context, a central problem is to investigate the gain that the extra information brings to the detection problem, compared to the case where only edge observation is available.

To get theoretical insights into such a problem, it is often assumed that the graph is generated from a simple probabilistic model called Stochastic Block Model (SBM), in which the probability of edge existence is higher within the community than that between different communities [4]. For the sole presence of SBM, as the size of community grows, the error rate of many algorithms decreases to zero in both the exact recovery and weak recovery metric [5], [6]. For the special case of a two-community model, the optimal error rate for weak recovery has been obtained as $n^{-(\sqrt{a}-\sqrt{b})^2/2}$ where a, b are parameters of SBM [7].

With the presence of extra node information, the condition of exact recovery is improved and generalized [8]. However, previous study does not exactly quantify the optimal error rate of SBM with side information. This paper will fill the gap by considering a model of two-community SBM with extra node feature vectors. We have obtained that the exact recovery error decreases polynomially in a rate quantified by $\gamma D_{1/2}(p_0||p_1) + (\sqrt{a} - \sqrt{b})^2 - 2$. In this expression, the contribution of side information to the error rate is coded in Rényi divergence. The optimal error rate on the extended model is achieved by maximum likelihood method, which is theoretically justified but can not be applied directly in practical problems without approximation. For many other implementable algorithms like variants of SDP relaxation and spectral clustering, their error rate decreases to zero but may not achieve the fundamental limit given in this paper. Nevertheless, the study of the optimal error rate provides a unified way to compare different algorithms in the experiment level.

This paper is organized as follows. In Section II, we review the previous works which are related with ours. In Section III, we introduce the mathematical model. Then in the following two sections, we present our error rate results for two different parameter regimes respectively. Finally the article concludes in Section VI.

The following notations are used throughout this paper: the random undirected graph G is written as $G(V, E)$ with vertex set V and edge set E; $V = \{1, \dots, n\} =: [n]$; $\mathcal{X}$ is the alphabet set of the random variable X; m is the number of samples generated at each node; $\mathrm{Bern}(p)$ and $\mathrm{Binom}(n, p)$ represent Bernoulli and Binomial distribution respectively; $f(n) = \omega(g(n))$ (or $= o(g(n))$) means that $\lim_{n\to\infty} f(n)/g(n) = \infty$ (or $= 0$); $\mathbb{1}[A]$ is the indicator function for the event A; W^n is the n-ary Cartesian power of the set W; The Hamming distance of two n-dimensional vectors is written as $\mathrm{dist}(x, y) := \sum_{i=1}^n \mathbb{1}[x_n \neq y_n]$ for $x, y \in \{\pm 1\}^n$.

II. Related Works

The model considered in this work extends the two-community SBM in [9]. Specifically, we assume the extra feature vectors of each node are independent samples, whose distribution depends on the label of the node. This model has been studied in Section V-B of [8], in which the number of features m is required to be of the order $\log n$ for side information to take effects. A general case of side information is studied in [10] and the exact recovery condition is obtained, which involves an optimization problem. We emphasize that the SBM in Theorem 4 of [10] assumes that the node labels are independently generated from $\mathrm{Bern}(\frac{1}{2})$ while the model in this paper requires uniform distribution over the space $\{Y_i \in \{\pm 1\} | \sum_{i=1}^n Y_i = 0\}$ where Y_i is the label of the i-th node. To distinguish the two models, we call the former

SBM with equal probability and the latter the SBM with equal community size.

In previous researches of SBM, the recovery condition is extensively studied, in which the error rate converges to zero [11]. For SBM model with side information, we find the error rate of SBM with equal community size constraint allows close-form solution in this paper while the error rate for SBM with equal probability remains an open problem.

Rényi divergence has been used in SBM in [7] to characterize the optimal error rate in weak recovery sense. In that study, both the dense and sparse graph are considered. In this paper, we consider the optimal error rate in exact recovery metric and obtain similar results containing Rényi divergence in both the two types of graphs for SBM with side information.

III. Mathematical Models

The two-community symmetric stochastic block model (SSBM) is a special case of SBM, and we give the formal definition of SSBM as follows:

Definition 1 (SSBM)**.** *Let $0 \leq q < p \leq 1$ and $V = [n]$. The random vector $Y = (Y_1, \ldots, Y_n) \in \{\pm 1\}^n$ and random graph G are drawn under $\mathrm{SSBM}(n, p, q)$ if*

1) *Y is drawn uniformly with the constraint that $Y_1 + \cdots + Y_n = 0$ for $Y_i \in \{\pm 1\}$;*
2) *There is an edge of G between the vertices i and j with probability p if $Y_i = Y_j$ and with probability q if $Y_i \neq Y_j$; the existence of each edge is independent with each other.*

Sampling from SSBM, we can get a pair (Y, G) where each feasible label $Y = y$ has probability $1/\binom{n}{n/2}$. Within this probabilistic setting, the community detection task is to infer Y from G. When additional node observations X are added, we expect that better inference accuracy of Y is achieved using (G, X). The additional node observation is called side information, and we define it formally in the following model:

Definition 2 (SBMSI)**.** *Let (Y, G) be sampled from $\mathrm{SSBM}(n, p, q)$, and $X_{i1}, \ldots, X_{im}$ are i.i.d. random variables for $i \in [n]$, whose probability density function $p(x)$ is determined by Y_i as*

$$p(x) = \begin{cases} p_0(x) & Y_i = 1 \\ p_1(x) & Y_i = -1 \end{cases} \tag{1}$$

We call the above generative model as SSBM with symmetric side information (SBMSI) with parameter (n, m, p, q, p_0, p_1).

The node observations can be written concisely as $\{X_{ij} | i \in [n], j \in [m]\}$. Besides, the graph G can be regarded as observations of edges, and we can denote the edge observations in a similar way by using $Z_{ij} := \mathbb{1}[\{i, j\} \in E(G)]$. Using X_{ij} and Z_{ij}, the likelihood function for given Y is

$$p(x, z | Y = y) = p(z|y) \prod_{i=1}^{n} \prod_{j=1}^{m} p_0^{\sigma_i}(x_{ij}) p_1^{1-\sigma_i}(x_{ij}) \tag{2}$$

where $p(z|y)$ is the likelihood function for SSBM and $\sigma_i = (1 + y_i)/2$. Based on (2), we can use the maximum likelihood (ML) method to estimate Y:

$$\begin{aligned} \hat{Y} &= \arg\max_{y} \ p(x, z | Y = y) \\ &\text{s.t. } y_i \in \{\pm 1\}, \sum_{i=1}^{n} y_i = 0 \end{aligned} \tag{3}$$

The estimator $\hat{Y}$, given by (3), is an ML estimator in restricted parameter space. In contract, ML estimator for $y \in \{\pm 1\}^n$ (unrestricted parameter space) should be used for SBM with equal probability. To study the performance of the ML estimator, we need a metric of the error rate, whose formal definition is given as:

Definition 3 (Error Rate of Exact Recovery for SBMSI)**.** *Let (Y, Z, X) be sampled from $\mathrm{SBMSI}(n, m, p, q, p_0, p_1)$. For an algorithm that takes (Z, X) as inputs and outputs $\hat{Y} = \hat{Y}(Z, X)$, we define its error rate of exact recovery as $P_e := P(\hat{Y} \neq Y)$.*

The above definition is slightly different from that of SBM as the latter uses $\hat{Y} \neq \pm Y$. When no side information is available, we can only expect a recovery up to a global sign. However, since $p_0 \neq p_1$, the sign of Y can also be determined when side information is in hand.

We make another remark that the exact recovery metric imposes stricter requirement on the recovery algorithm than its weak recovery counterpart, which uses $\mathbb{E}[\mathrm{dist}(\hat{Y}, Y)]/n$ as the error rate.

Below we analyze the exact recovery error of the maximum likelihood estimator $\hat{Y}$ given by (3). Without edge observations, the estimation is decomposed into n independent hypothesis testing problems with the global constraint $\sum_{i=1}^{n} Y_i = 0$. In such case, Rényi divergence with order $\frac{1}{2}$ is used to quantify the error exponent [12]. This information theoretic quantity can be written as:

$$D_{1/2}(p_0 || p_1) = -2 \log\Big(\sum_{x \in \mathcal{X}} \sqrt{p_0(x) p_1(x)}\Big) \tag{4}$$

With node observations, we divide our discussion into two cases:

1) dense SBMSI: p, q are constant values;
2) sparse SBMSI: $p = a \log n / n, q = b \log n / n$ and a, b are constant values.

The recovery error rate for the first case is given in Section IV while the latter case is analyzed in Section V.

IV. Error Exponent for Dense SBMSI

Theorem 1. *Let $\gamma = \frac{m}{n}$ be a constant. If p, q are constant, using maximum likelihood estimator* (3)*, the error exponent of exact recovery is given by:*

$$-\lim_{n \to \infty} \frac{1}{n} \log P_e = \gamma D_{1/2}(p_0 || p_1) + D_{1/2}(\mathrm{Bern}(p) || \mathrm{Bern}(q)) \tag{5}$$

From Theorem 1, we see that the recovery error decreases in exponential rate. When $\gamma = 0$, Theorem 1 says $D_{1/2}(\mathrm{Bern}(p)||\,\mathrm{Bern}(q))$ is the error exponent for exact recovery. Since weak recovery differs from exact recovery by a polynomial factor, $D_{1/2}(\mathrm{Bern}(p)||\,\mathrm{Bern}(q))$ is also the exponent for weak recovery, which has been obtained in [7]. Besides, assuming γ is an integer, the error exponent can be regarded as the Rényi divergence between the joint distribution $\underbrace{p_0 \times \cdots \times p_0}_{\gamma} \times \mathrm{Bern}(p)$ and $\underbrace{p_1 \times \cdots \times p_1}_{\gamma} \times \mathrm{Bern}(q)$. By independent conditions, we can decompose this divergence in the summation form. Furthermore, the result of Theorem 1 requires m and n have the same order. When $m = o(n)$, the edge information takes dominate effects; when $m = \omega(n)$, the side information dominates and the edge information is negligible.

A. Proof of Theorem 1

We introduce some additional notations used throughout this proof. Let $|A|, A^c$ be the cardinality, complement of the set A, respectively. For distributions p_0 and p_1, $D(p_0||p_1)$ is the Kullback-Leibler divergence. Let $p_{B_q}(z) = q^z(1-q)^{1-z}$ be the probability mass distribution for $\mathrm{Bern}(q)$. From type theory, the set of possible types for m samples with alphabet $\mathcal{X}$ is denoted as $\mathcal{P}_m$. For any $P \in \mathcal{P}_m$, the probability of the type class $T(P)$ under distribution p_i is denoted as $Q_i^m(T(P))$.

Let us consider $P(\hat{Y} = Y|Y = y^*)$ for a certain labeling of nodes y^*. Since Y is uniformly sampled, $P_e = P(\hat{Y} = Y|Y = y^*)$. If the ML in (3) fails to exactly recover y^*, then there exists $y \neq y^*$ such that $p(x, z|y) > p(x, z|y^*)$. Let F_k denote the event when there are k pairs differences between y and y^*.

$$F_k := \{\exists y \in \{\pm 1\}^n |\, \mathrm{dist}(y, y^*) = 2k, p(x, z|y) > p(x, z|y^*)\} \tag{6}$$

Since y is expected to satisfy the constraint $\sum_{i=1}^n y_i = 0$, $\mathrm{dist}(y, y^*)$ is only allowed to take even values. Taking logarithm on both sides of $p(x, z|y) > p(x, z|y^*)$, we get the equivalent inequality:

$$\sum_{i=1}^{km} \left(\log \frac{p_1(x_{1i})}{p_0(x_{1i})} + \log \frac{p_0(x_{2i})}{p_1(x_{2i})} \right) \geq \log \frac{p(1-q)}{q(1-p)} \sum_{i=1}^{k(n-2k)} (z_i - z_i') \tag{7}$$

where $x_{1i}(x_{2i})$ are sampled from $p_0(p_1)$ respectively, and $z_i \sim \mathrm{Bern}(p), z_i' \sim \mathrm{Bern}(q)$.

We denote the event described by (7) as A_k, and each F_k can be regarded as the union of $\binom{n/2}{k}^2$ events of A_k for different node indexes.

To obtain an upper bound of $P(A_k)$, we further define several empirical distributions as follows:

$$P(\widetilde{X}_j = u) = \frac{1}{km} \sum_{i=1}^{km} \mathbb{1}[x_{ji} = u] \text{ for } u \in \mathcal{X}, j = 1, 2$$

$$P(\widetilde{Z} = u) = \frac{1}{k(n-2k)} \sum_{i=1}^{k(n-2k)} \mathbb{1}[z_i = u], u \in \{0, 1\}$$

and $\widetilde{Z}'$ is defined similarly. Then (7) is transformed as

$$m \left[\sum_{x \in \mathcal{X}} P_{\widetilde{X}_1}(x) \log \frac{p_1(x)}{p_0(x)} + \sum_{x \in \mathcal{X}} P_{\widetilde{X}_2}(x) \log \frac{p_0(x)}{p_1(x)} \right] + (n-2k) \left[\sum_{z \in \{0,1\}} P_{\widetilde{Z}}(z) \log \frac{p_{B_q}(z)}{p_{B_p}(z)} + \sum_{z \in \{0,1\}} P_{\widetilde{Z}'}(z) \log \frac{p_{B_p}(z)}{p_{B_q}(z)} \right] \geq 0 \tag{8}$$

When p, q are constant, using (8), $P(A_k)$ can be estimated by Sanov's theorem: $-\frac{1}{kn} \log P(A_k) \to \theta_k^*$ as $n \to \infty$ where

$$\begin{aligned} \theta_k^* = & \min_{\widetilde{X}_1, \widetilde{X}_2, \widetilde{Z}, \widetilde{Z}'} \gamma(D(p_{\widetilde{X}_1}||p_0) + D(p_{\widetilde{X}_2}||p_1)) + \\ & (1 - \frac{2k}{n})(D(p_{\widetilde{Z}}||\,\mathrm{Bern}(p)) + D(p_{\widetilde{Z}'}||\,\mathrm{Bern}(q))) \\ & \text{s.t. } (\widetilde{X}_1, \widetilde{X}_2, \widetilde{Z}, \widetilde{Z}') \text{ satisfy (8)} \end{aligned}$$

Using the Lagrange multiplier, we can get

$$\begin{aligned} p_{\widetilde{X}_1}(x) &= c_1 p_0^{1-\lambda}(x) p_1^{\lambda}(x) \quad p_{\widetilde{X}_2}(x) = c_2 p_1^{1-\lambda}(x) p_0^{\lambda}(x) \\ p_{\widetilde{Z}}(z) &= c_3 p_{B_p}^{1-\lambda}(x) p_{B_q}^{\lambda}(z) \quad p_{\widetilde{Z}'}(z) = c_4 p_{B_q}^{1-\lambda}(x) p_{B_p}^{\lambda}(z) \end{aligned}$$

where $c_1, \ldots, c_4$ are normalization coefficients for these distributions. The parameter λ is chosen such that (8) becomes equality, which leads to $\lambda = \frac{1}{2}$. Therefore, $\theta_k^* = \gamma D_{1/2}(p_0||p_1) + (1 - \frac{2k}{n}) D_{1/2}(\mathrm{Bern}(p)||\,\mathrm{Bern}(q))$. Denoting $C_1 = \gamma D_{1/2}(p_0||p_1)$, $C_2 = D_{1/2}(\mathrm{Bern}(p)||\,\mathrm{Bern}(q))$ for short, then $P(A_k) \leq \exp(-knC_1 - k(n-2k)C_2)$ for large n. Using the union bound, we can control $P(F_k)$ by

$$P(F_k) \leq \binom{n/2}{k}^2 P(A_k) \tag{9}$$

and by $\binom{n}{k} \leq (ne/k)^k$, we can further bound P_e above as follows:

$$\begin{aligned} P_e &\leq \sum_{k=1}^{n/4} \binom{n/2}{k}^2 P(A_k) \\ &\leq \sum_{k=1}^{n/4} \exp(-nf(k)) \end{aligned}$$

where $f(k) = \frac{2k}{n} \log \frac{2k}{ne} + k(C_1 + C_2) - \frac{2k^2}{n} C_2$. By computing $f'(x) = \frac{2}{n} \log \frac{2x}{n} + C_1 + C_2 - \frac{4C_2 x}{n}$, $1 \leq x \leq \frac{n}{4}$. $f'(1) > 0, f'(\frac{n}{4}) > 0 \Rightarrow f'(x) > 0$ for $1 \leq x \leq \frac{n}{4}$. Therefore, $f(x)$ increases in the interval $[1, \frac{n}{4}]$, and $f(k) \geq f(1)$ for $1 \leq k \leq \frac{n}{4}$.

$$P_e \leq \frac{n}{4} \exp(-nf(1)) = \exp(-n(C_1 + C_2 + o(1))) \tag{10}$$

On the other hand, $P_e \geq P(A_1) = \exp(-n(C_1 + C_2 + o(1)))$. Finally we have $-\frac{1}{n} \lim_{n \to \infty} \log P_e = C_1 + C_2$, and the proof of Theorem 1 is completed.

V. Error Rate for Sparse SBMSI

In Section IV we discussed the exponential error rate for dense SBMSI. In this section, we will present the polynomial error rate for sparse SBMSI. This result is summarized in the following theorem:

Theorem 2. *Let $\gamma = \frac{m}{\log n}$ be a constant. If $p = a\log n/n$ and $q = b\log n/n$, using maximum likelihood estimator* (3)*, if*

$$\gamma D_{1/2}(p_0||p_1) + (\sqrt{a}-\sqrt{b})^2 - 2 > 0 \tag{11}$$

then the error probability of exact recovery is bounded by

$$P_e \le (\frac{1}{4}+o(1))n^{-\left(\gamma D_{1/2}(p_0||p_1)+(\sqrt{a}-\sqrt{b})^2-2+o(1)\right)} \tag{12}$$

If the following condition

$$(\sqrt{a}-\sqrt{b})^2 - 2 > 3a^{1/3}b^{1/3}(a^{1/6}-b^{1/6})^2 \tag{13}$$

is satisfied, then we can show that P_e is lower bounded by

$$P_e \ge (\frac{1}{4}+o(1))n^{-\left(\gamma D_{1/2}(p_0||p_1)+(\sqrt{a}-\sqrt{b})^2-2+o(1)\right)} \tag{14}$$

Theorem 2 tells us that the side information X increases the decay rate of error probability P_e quantified by $\gamma D_{1/2}(p_0||p_1)$. Under some parameter configurations specified in (13), the quantity $\gamma D_{1/2}(p_0||p_1) + (\sqrt{a}-\sqrt{b})^2 - 2$ exactly describes the error rate for the exact recovery problem of SBMSI.

We notice that when (13) is satisfied, so is (11). In such case, the error rate is given by

$$-\lim_{n\to\infty}\frac{\log P_e}{\log n} = \gamma D_{1/2}(p_0||p_1) + (\sqrt{a}-\sqrt{b})^2 - 2$$

To obtain this error rate, we need a slightly stronger condition (13) than the recovery threshold $\sqrt{a}-\sqrt{b} > \sqrt{2}$ for SSBM.

In addition, when $p_0 = p_1$, Theorem 2 gives the error rate of maximum likelihood for SSBM. This corollary is summarized as follows:

Corollary 1. *Consider* $\mathrm{SSBM}(n, \frac{a\log n}{n}, \frac{b\log n}{n})$ *with equal community size. For ML algorithms, the exact recovery error rate P_e satisfies*

$$\lim_{n\to\infty}\frac{\log P_e}{\log n} = 2 - (\sqrt{a}-\sqrt{b})^2 \tag{15}$$

as long as (13) *holds.*

A. Proof of Theorem 2

We start from (8) to get the upper and lower bounds for P_e. Firstly, we introduce the following lemma, which gives the lower bound of $P(A_1)$ when $p, q = O(\frac{\log n}{n})$.

Lemma 1. *For event A_1 specified in* (7) *with $k = 1$, we have the following estimation*

$$P(A_1) \ge \exp(-(\gamma D_{1/2}(p_0||p_1) + (\sqrt{a}-\sqrt{b})^2 + o(1))\log n) \tag{16}$$

Proof of Lemma 1. When $k = 1$, the inequality (7) can be rewritten as $\sum_{i=1}^{n-2}(z_i' - z_i) \ge \epsilon$ where

$$\epsilon := \frac{m}{\log a/b}\cdot\Big[D(P_{\widetilde{X}^1}||P_1) - D(P_{\widetilde{X}^1}||P_0) + D(P_{\widetilde{X}^2}||P_0) - D(P_{\widetilde{X}^2}||P_1)\Big],$$

Let $P_{\widetilde{X}^{i_1}}$ and $P_{\widetilde{X}^{i_2}}$ follow the distribution $P(X=x) = \frac{\sqrt{p_0(x)p_1(x)}}{\sum_{x\in\mathcal{X}}\sqrt{p_0(x)p_1(x)}}$, which makes $\epsilon = 0$. For this special choice of distribution $P_{\widetilde{X}^{i_1}}$ and $P_{\widetilde{X}^{i_2}}$, using Sanov's theorem, we have that

$$\begin{aligned}P(A_1) &\ge \frac{1}{(m+1)^{2|\mathcal{X}|}}\exp(-m(D(p_{\widetilde{X}_1}||p_0) + D(p_{\widetilde{X}_2}||p_1))\\ &\cdot P\left(\sum_{i=1}^{n-2}(z_i'-z_i)\ge 0\right)\\ &= \exp(-\log n(\gamma D_{1/2}(P_0||P_1)+o(1)))P(\sum_{i=1}^{n-2}(z_i'-z_i)\ge 0).\end{aligned}$$

From Lemma 4 from [11], $P(\sum_{i=1}^{n-2}(z_i'-z_i)\ge 0)$ is lower bounded by $n^{-(\sqrt{a}-\sqrt{b})^2+o(1)}$. Therefore, (16) is obtained. □

Lemma 2. *Let p_0, p_1 be two probability distributions defined on alphabet $\mathcal{X}$, then the following inequality holds*

$$\left(\sum_{x\in\mathcal{X}} p_0^{\frac{1}{3}}(x)p_1^{\frac{2}{3}}(x)\right)^3 \le \left(\sum_{x\in\mathcal{X}}\sqrt{p_0(x)p_1(x)}\right)^2 \tag{17}$$

Proof. Let $f(x) = p_0^{\frac{1}{3}}(x)p_1^{\frac{1}{3}}(x)$, $g(x) = p_1^{\frac{1}{3}}(x)$, $p = \frac{3}{2}, q = 3$. We can verify $\frac{1}{p}+\frac{1}{q} = 1$. By Hölder's inequality:

$$(\sum_{x\in\mathcal{X}} f(x)g(x)) \le (\sum_{x\in\mathcal{X}} f^p(x))^{\frac{1}{p}}(\sum_{x\in\mathcal{X}} g^q(x))^{\frac{1}{q}} \tag{18}$$

Since $\sum_{x\in\mathcal{X}} p_1(x) = 1$, (18) implies (17). □

Proof of Theorem 2. Below we use Chernoff's inequality to give an upper bound of $P(A_k)$: $P(A_k) \le n^{-k\theta_k^*}$ where $\theta_k^* = \gamma D_{1/2}(p_0||p_1) + (1-\frac{2k}{n})(\sqrt{a}-\sqrt{b})^2$.

$$\begin{aligned}P(A_k) &\le \mathbb{E}\left[\exp\left(s\sum_{i=1}^{km}\left(\log\frac{p_1(x_{1i})}{p_0(x_{2i})}+\log\frac{p_0(x_{2i})}{p_1(x_{2i})}\right)\right)\right]\\ &\cdot\mathbb{E}\left[\exp\left(s\log\frac{a}{b}\sum_{i=1}^{k(n-2k)}(z_i'-z_i)\right)\right]\\ &\overset{(a)}{=} (\sum_{x\in\mathcal{X}} p_0^{1-s}(x)p_1^s(x))^{km}(\sum_{x\in\mathcal{X}} p_1^{1-s}(x)p_0^s(x))^{km}\\ &\cdot\exp(k\log n(1-\frac{2k}{n})(-a-b+a^sb^{1-s}+b^sa^{1-s}+o(1)))\end{aligned}$$

where (a) follows from independence condition. Choosing $s = \frac{1}{2}$, we then have $P(A_k) \le n^{-k(\theta_k^*+o(1))}$.

When $k \ge \frac{n}{\sqrt{\log n}}$, using Lemma 8 of [13], $P(F_k)$ decreases exponentially. The error probability for $2 \le k < \frac{n}{\sqrt{\log n}}$ is analyzed using (9).

$$\begin{aligned}P_e &\le P(F_1) + (1+o(1))\sum_{k=2}^{\frac{n}{\sqrt{\log n}}} P(F_k) \le (1+o(1))\\ &\cdot\sum_{k=2}^{\frac{n}{\sqrt{\log n}}}\exp(k(-\mu\log n + \frac{2k}{n}\log n(\sqrt{a}-\sqrt{b})^2 - 2\log 2k + 2))\end{aligned}$$

where μ is defined as

$$\mu = (\sqrt{a} - \sqrt{b})^2 - 2 + \gamma D_{1/2}(p_0||p_1) > 0 \qquad (19)$$

For $P(F_1)$, we have $P(F_1) \leq (n/2)^2 P(A_1) \leq \frac{1}{4} n^{-\mu+o(1)}$. For $2 \leq k \leq \frac{n}{\sqrt{\log n}}$, using the inequality

$$\frac{2k}{n}(\sqrt{a} - \sqrt{b})^2 \log n - 2\log 2k + 2 \leq C\sqrt{\log n}$$

we can obtain

$$\begin{aligned} P_e &\leq \frac{1}{4} n^{-\mu+o(1)} + (1+o(1)) \sum_{k=2}^{\frac{n}{\sqrt{\log n}}} \exp(k((-\mu+o(1))\log n)) \\ &= \frac{1}{4} n^{-\mu+o(1)} + (1+o(1)) \frac{n^{-\mu+o(1)}}{1 - n^{-\mu+o(1)}} \\ &= (\frac{1}{4} + o(1)) n^{-\mu+o(1)} \end{aligned}$$

Therefore, (12) is established.

To prove the lower bound (14), we first define the event A_{ij} as $p(x,z|y) > p(x,z|y^*)$ where y is defined as $y_s = -y_s^*, s = i,j$ and $y_s = y_s^*$ otherwise.

Define $S := \{(i,j)|1 \leq i < j \leq n, y_i^* = -y_j^*\}$. Then it follows that $\cup_{(i,j)\in S} A_{ij} \subset F$, where F denotes the event when ML fails to recover the community labels exactly. By Bonferroni inequality,

$$P(\bigcup_{(i,j)\in S} A_{ij}) \geq \sum_{(i,j)\in S} P(A_{ij}) - \sum_{(i,j)\neq(r,s)} P(A_{ij} \cap A_{rs}) \qquad (20)$$

To get the lower bound of $P(\cup_{(i,j)\in S} A_{ij})$ by (20), we need a lower bound for $\sum_{(i,j)\in S} P(A_{ij})$ and an upper bound for $\sum_{(i,j)\neq(r,s)} P(A_{ij} \cap A_{rs})$. We first deal with $P(A_{ij})$. Notice that the single event A_{ij} is equivalent with A_1. By Lemma 1, $P(A_{ij})$ is lower bounded by $n^{-\gamma D_{1/2}(p_0||p_1) - (\sqrt{a}-\sqrt{b})^2 + o(1)}$. Since $|S| = (n/2)^2$, the term $\sum_{(i,j)\in S} P(A_{ij})$ is of order $\frac{1}{4} n^{-\mu+o(1)}$. Next we give the upper bound of $P(A_{ij} \cap A_{rs})$ according to two cases.

First is the case when $|\{i,j,r,s\}| = 4$. Then $A_{ij} \cap A_{rs}$ implies the event $A_{ijrs} : p(x,z|y^{(1)})p(x,z|y^{(2)}) > p^2(x,z|y^*)$ where $y^{(1)}$ differs from y^* at position (i,j) and $y^{(2)}$ differs from y^* at position (r,s). After taking the logarithm on both sides and simplification, the inequality representation for the event A_{ijrs} is the same with A_2. Therefore, $P(A_{ij} \cap A_{rs}) \leq n^{-2(\theta_2^* + o(1))}$. The number of elements in the set $S_1 := \{(i,j,r,s)|i<j, r<s, |\{i,j,r,s\}| = 4\}$ is $\binom{n}{4} \leq n^4$. Therefore, the probability sum $\sum_{(i,j,r,s)\in S_1} P(A_{ij} \cap A_{rs}) \leq n^{-2\mu+o(1)}$, which has smaller order than $n^{-\mu+o(1)}$ since $\mu > 0$.

Another case happens when $|\{i,j,r,s\}| = 3$. Under such case, without loss of generality we can assume $i = r, y_i^{(1)} = y_r^{(2)} = 1$. For the case $y_i^{(1)} = y_r^{(2)} = -1$, we only need to exchange p_0 and p_1, and the following analysis is still valid. Then

$$A_{ijrs} : 2\sum_{i=1}^{m} \log \frac{p_1(x_{1i})}{p_0(x_{2i})} + \sum_{i=1}^{2m} \log \frac{p_0(x_{2i})}{p_1(x_{2i})}$$

$$+ \log \frac{a}{b} \left(\sum_{i=1}^{n} (z_i' - z_i) + 2 \sum_{i=n+1}^{3n/2} (z_i' - z_i) \right) \geq 0$$

Using Chernoff's inequality, we can write an upper bound of $P(A_{ijrs})$ as

$$\begin{aligned} P(A_{ijrs}) \leq & \left(\sum_{x\in\mathcal{X}} p_0^{1-2s} p_1^{2s} \right)^m \left(\sum_{x\in\mathcal{X}} p_1^{1-s} p_0^s \right)^{2m} \\ & \cdot \exp\Big(\log n (-\frac{3}{2}(a+b) + a\exp(-s\log\frac{a}{b}) + b\exp(s\log\frac{a}{b}) \\ & + \frac{a}{2}\exp(-2s\log\frac{a}{b}) + \frac{b}{2}\exp(2s\log\frac{a}{b}) + o(1)) \Big) \end{aligned}$$

Let $s = \frac{1}{3}$. We then have

$$\begin{aligned} P(A_{ijrs}) &\leq (\sum_{x\in\mathcal{X}} p_0^{1/3}(x) p_1^{2/3}(x))^{3m} \\ &\quad \cdot \exp(\frac{3}{2}\log n(-a-b+a^{1/3}b^{2/3} + a^{2/3}b^{1/3} + o(1))) \\ &\leq \exp(-\log n(\gamma D_{1/2}(p_1||p_0) \\ &\quad + \frac{3}{2}(a+b-a^{1/3}b^{2/3} - a^{2/3}b^{1/3}) + o(1))) \end{aligned}$$

where the last inequality follows from Lemma 2. It then follows that

$$P(A_{ijrs}) \leq n^{-\mu'/2 - 1 - (\gamma D_{1/2}(p_0||p_1) + (\sqrt{a}-\sqrt{b})^2) + o(1)}$$

where $\mu' = (\sqrt{a} - \sqrt{b})^2 - 2 - 3a^{1/3}b^{1/3}(a^{1/6} - b^{1/6})^2 > 0$ from (13). The set $S_2 := \{(i,j,r,s)|i < j, r < s, |\{i,j,r,s\}| = 3\}$ has at most n^3 such terms, then we have $\sum_{(i,j,r,s)\in S_2} P(A_{ij} \cap A_{rs}) \leq n^{-\mu'/2 - \mu + o(1)}$, which has smaller order than $n^{-\mu+o(1)}$.

Based on the above discussion, $\sum_{(i,j,r,s)\in S_2} P(A_{ij} \cap A_{rs}) \leq n^{-2\mu+o(1)} + n^{-\mu-\mu'/2} = o(1) n^{-\mu+o(1)}$. Then we conclude that

$$\begin{aligned} P_e = P(F) &\geq P(\cup_{(i,j)\in S} A_{ij}) \\ &\geq \frac{1}{4} n^{-\mu+o(1)} - o(1) n^{-\mu+o(1)}, \text{ from (20)} \\ &= (\frac{1}{4} + o(1)) n^{-\mu+o(1)} \end{aligned}$$

□

VI. Conclusion

In this paper, we obtain the optimal error rate in the sense of exact recovery for a two-community SBM with side information. Our result shows that the detection error can be characterized by Rényi divergence and the parameters of SBM. To control the recovery error within a given level, our result shares insight on the necessary number of features and nodes. Whether the condition (13) for the error rate can be relaxed will be considered in the future study.

REFERENCES

[1] X. Dong, D. Thanou, L. Toni, M. Bronstein, and P. Frossard, "Graph signal processing for machine learning: A review and new perspectives," *IEEE Signal Processing Magazine*, vol. 37, no. 6, pp. 117–127, 2020.

[2] E. Mossel and J. Xu, "Local algorithms for block models with side information," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 2016, pp. 71–80.

[3] Y. Zhang, E. Levina, J. Zhu *et al.*, "Community detection in networks with node features," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3153–3178, 2016.

[4] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[5] S.-Y. Yun and A. Proutiere, "Accurate community detection in the stochastic block model via spectral algorithms," *arXiv preprint arXiv:1412.7335*, 2014.

[6] Y. Fei and Y. Chen, "Achieving the bayes error rate in stochastic block model by sdp, robustly," in *Conference on Learning Theory*. PMLR, 2019, pp. 1235–1269.

[7] A. Y. Zhang and H. H. Zhou, "Minimax rates of community detection in stochastic block models," *Ann. Statist.*, vol. 44, no. 5, pp. 2252–2280, 10 2016. [Online]. Available: https://doi.org/10.1214/15-AOS1428

[8] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.

[9] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, 2015, pp. 670–688.

[10] A. R. Asadi, E. Abbe, and S. Verdú, "Compressing data on graphs with clusters," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1583–1587.

[11] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[12] C. Gao, Z. Ma, A. Y. Zhang, H. H. Zhou *et al.*, "Community detection in degree-corrected block models," *Annals of Statistics*, vol. 46, no. 5, pp. 2153–2185, 2018.

[13] F. Zhao, M. Ye, and S.-L. Huang, "Exact recovery of stochastic block model by ising model," *Entropy*, vol. 23, no. 65, 2021.